Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients

Duy Dinh, Lynda Tamine

Vers un modèle d'indexation sémantique adapté aux dossiers médicaux de patients

Duy Dinh, Lynda Tamine

Université Paul Sabatier
118 route de Narbonne
31062 Toulouse - Cedex 9
dinh@irit.fr, lechani@irit.fr

ABSTRACT. *This paper presents a semantic model adapted for the indexing of electronic patient records (EHRs) as a support to the process of finding medical information. Given the specificity of such documents, the indexing is based on the sequence of Natural Language Processing steps: semantic annotation based on the use of the MeSH (Medical Subject Headings) thesaurus, concept disambiguation, extraction of clinical values, and concept weighting. The weighting scheme takes into account the granularity level of the index (patient documents or patient records) and the location of concepts in documents and in the MeSH hierarchy in order to translate both their specificity and their centrality. The proposed indexing model is evaluated on a corpus of EHRs and shows its effectiveness for indexing such documents.*

RÉSUMÉ. *Ce papier présente un modèle d'indexation sémantique adapté aux dossiers électroniques de patients. Ce modèle servira de support à des processus de recherche d'information médicale, permettant à terme de promouvoir l'expérience collective des médecins. Compte tenu de la spécificité de ce type de documents, le processus d'indexation est basé sur la succession d'étapes d'annotation sémantique fondée sur l'utilisation de MeSH (Medical Subject Headings), de désambiguïsation répondant au problème d'homonymie, d'extraction de valeurs cliniques, puis de pondération des concepts. Le schéma de pondération tient compte du niveau de description de l'index (document ou dossier) ainsi que de la localisation des concepts dans le document et dans la hiérarchie de MeSH et ce, dans le but de traduire à la fois leur spécificité et leur centralité. Le modèle d'indexation proposé est évalué sur un corpus de dossiers électroniques de patients et montre son efficacité pour ce type de documents.*

KEYWORDS: *Semantic Indexing, Information Retrieval, Information Extraction, Medical Informatics, Patient Records.*

MOTS-CLÉS : *Indexation Sémantique, Recherche d'Information, Informatique médicale, Extraction d'Information, Dossier médical de Patients.*

1. Introduction

Les systèmes d'information médicaux ont connu une grande évolution depuis ces deux dernières décennies tant du point de vue de leur architecture que de la qualité et de la diversité des services autour du stockage de l'information, l'accès à l'information pertinente pour une médecine basée sur des niveaux de preuve, l'aide à la décision pour l'amélioration de la qualité des soins (Hersh, 2008).

Dans ce cadre général, l'information biomédicale utilisée comme support pour les tâches de recherche, d'extraction d'information et de connaissances (datamining), concerne principalement la littérature médicale et les dossiers médicaux de patients qui font l'objet de notre travail décrit dans le présent papier. Ces dossiers, comprenant des éléments d'identification d'ordre administratif et des données médicales, constituent un noyau fondamental de la qualité des soins dans les structures hospitalières. En France, la loi du 4 mars 2002 a conféré au dossier médical du patient ou plus connu sous le nom du dossier médical personnalisé (DMP), particulièrement à son accessibilité et à la maîtrise de ses flots d'information, une importance capitale. Une revue de l'état de l'art autour de l'exploitation de ces ressources d'information médicales pour l'amélioration des services de soins de manière générale (Hersh, 2008), (Friedman, 2005), (Uzuner *et al.*, 2008) révèle un constat majeur : en pratique, ces dossiers sont constitués de collections de documents très hétérogènes du point de vue du format, du contenu et de la sémantique :

(a) du format : texte semi-structuré ou non structuré comme les comptes rendus d'intervention, de consultation, d'anatomopathologie, d'images telles que les PET (Positron à Emission de Tomographie), images de résonance magnétique etc.

(b) du contenu : texte en langage naturel qui exprime des faits observables, des décisions médicales de prévention, de diagnostique et de traitement, ne suivant pas de régularités grammaticales ou lexicales et présentant des valeurs structurelles enfouies dans le texte (données démographiques, mesures d'analyse telles que le taux de cholestérol, le taux de glucose, la pression artérielle, etc.).

(c) de la sémantique : l'intelligibilité d'un document est dépendante de l'historique thérapeutique du patient contenu dans d'autres documents du même patient.

Ce travail porte sur l'indexation automatique de documents textuels contenus dans les DMP, comme étape clé, en amont à toute exploitation du dossier : recherche d'information, catégorisation pathologique, assistance à la concertation médicale etc.

A notre connaissance, les travaux de recherche portant sur l'indexation de ressources d'information médicales ont essentiellement porté sur la littérature scientifique sous forme de résumés d'articles, plus rarement de textes intégraux (Névéol *et al.*, 2006), (Aronson, 2001) en utilisant des ressources terminologiques comme

UMLS [1], MeSH [2] et SNOMED [3], GO [4], Biothesaurus (Liu *et al.,* 2006) ou Drugbank (Kolářik *et al.,* 2007). Dans ce cadre, des modèles d'indexation sémantiques semi-automatiques de textes ont été proposés selon deux principales approches : (a) linguistique fondée sur l'analyse lexicale et syntaxique de la structure du discours, (b) statistique fondée sur la distribution des concepts dans les documents.

Concernant l'indexation de DMP, la revue de la littérature révèle que les travaux sont moins abondants (Hersh, 2008). Les DMP ont été essentiellement abordés sous l'angle de l'identification de catégories sémantiques pertinentes et statuts associés retrouvés dans les résumés de sortie de l'hôpital (Subramaniam *et al.,* 2003), (Long, 2007), (Uzuner *et al.,* 2008), de la détection de relations sémantiques (Sibanda, 2006).

L'objectif de notre travail est de proposer un processus et modèle d'indexation dédiés à un document spécialisé en l'occurrence le DMP. Pour cela, nous répondons particulièrement à deux questions :

1) Quelles sont les étapes du processus d'indexation permettant d'identifier les concepts non-ambigus dans les documents ainsi que les valeurs associées ?

2) Quel est le modèle d'index à générer qui considère la distribution des concepts dans les documents et dans les DMP en qualité d'unités d'information accessibles?

La suite de cet article est organisée comme suit : La section 2 présente un état de l'art sur l'indexation des sources d'information médicales puis positionne notre contribution dans ce cadre. La section 3 décrit le processus d'indexation des DMP puis détaille le modèle d'indexation. Une évaluation expérimentale ainsi que les résultats sont présentés et discutés dans la section 4. La section 5 conclut le papier.

2. Indexation de documents médicaux : littérature biomédicale *vs.* dossiers médicaux de patients

On retrouve dans (Hersh, 2008) une classification des documents médicaux selon deux catégories :

(a) Connaissance du domaine médical : comprend les documents qui rapportent les publications scientifiques du domaine sous forme d'articles de journaux, d'ouvrages, de rapports ou de guides de bonnes pratiques.

(b) Informations spécifiques aux patients : comprennent des données structurées, semi-structurées ou narratives portant sur des faits observables, données factuelles, historique des patients, décisions médicales contenues dans les comptes-rendus (CR) de consultation, CR d'anatomie pathologique, CR opératoire, CR d'imagerie etc.

1. Unified Medical Language System
2. MEdical Subject Headings
3. Systematized NOmenclature of MEDicine
4. Gene Ontology

Ces deux types de documents sont différents du point de vue du contenu et par conséquent de l'usage. Les documents de la littérature sont caractérisés par un contenu homogène avec des sujets spécifiques et ont fait l'objet de nombreux travaux sur l'indexation et la recherche d'information (Avillach *et al.*, 2007), (Névéol *et al.*, 2007), (Pereira *et al.*, 2008). En revanche, le DMP est une collection de documents constituant un enregistrement électronique longitudinal d'information sur la santé des patients comprenant des faits, des décisions, des valeurs etc. Les travaux sur la représentation de l'information contenue dans le DMP sont plus ciblés (Price *et al.*, 2002), (Sibanda, 2006) et focalisent sur des tâches d'extraction d'information et de relations entre informations/connaissances.

Nous synthétisons dans cette section les travaux de représentation et/ou indexation ayant porté sur ces deux types de documents.

2.1. *Indexation de la littérature biomédicale*

Les volumes d'information dans la littérature biomédicale sont continuellement croissants. Leur accessibilité et indexation sont particulièrement confrontées aux problèmes de synonymie, homonymie et présence d'acronymes (Hersh, 2008). Il existe deux principales approches d'indexation qui sont l'approche *manuelle* et *(semi)-automatique*. L'indexation manuelle a été initiée par la NLM (National Library of Medicine) et a essentiellement servi à l'association de descripteurs sémantiques aux résumés des documents du corpus MEDLINE. L'accroissement des publications dans MEDLINE (en moyenne 2500 par jour) a conduit au développement d'outils d'indexation (semi) automatique tel que MTI (Aronson, 2001). L'indexation (semi)-automatique a été plus largement utilisée dans le domaine et vue comme une recommandation automatique de descripteurs (Kim *et al.*, 2001), (Cai *et al.*, 2004) ou de couples de descripteurs/qualificatifs (Névéol *et al.*, 2007). Concernant particulièrement l'indexation des ressources francophones, les travaux intègrent une ou plusieurs terminologies médicales (MeSH, ICD-10, CCAP, TUV) en associant des descripteurs MeSH aux documents dans le catalogue CiSMEF (Névéol *et al.*, 2006), (Pereira *et al.*, 2008).

Les descripteurs de l'index peuvent être ambigus. A titre d'exemple, une étude (Schiemann *et al.*, 2008) a montré qu'il existe 175 termes désignant à la fois des espèces et des protéines, 67 termes désignant des médicaments et des protéines, 123 termes désignant des cellules et des tissus. Des méthodes de désambiguïsation (WSD) ont été proposées pour associer automatiquement aux concepts leur sens "correct", équivalent à ceux qu'aurait annoté un expert. De nombreux travaux ont traité l'ambiguïté des termes issus d'UMLS (Widdows *et al.*, 2003), de MeSH (Névéol *et al.*, 2006), des acronymes (Gaudan *et al.*, 2005) et expressions de gènes (Andreopoulos *et al.*, 2008) en proposant des méthodes basées sur les distributions des fréquences des termes dans les sources ou en entraînant les différents sens des termes ambigus en se basant sur des méthodes d'apprentissage automatique ou de classification.

2.2. *Traitement automatique des dossiers médicaux de patients*

Le contenu des DMP comprend du texte écrit sous forme du langage naturel qui est jusqu'à présent le principal moyen de transmettre de l'information clinique entre les établissements de santé. De nombreux travaux dans le domaine ont été réalisés pour résoudre des problèmes spécifiques liés aux DMP et ce pour diverses applications : informatique décisionnelle médicale, surveillance des maladies infectieuses, codification automatique des pathologies, etc. Plus précisément, ces travaux portent essentiellement sur le traitement du texte dans les DMP pour les tâches suivantes :

– *Reconnaissance de catégories sémantiques* : les algorithmes utilisés pour cette tâche sont basés sur l'apprentissage supervisé avec les machines à vecteur de support (Sibanda, 2006) ou de patrons d'extraction utilisant des grammaires (Friedman, 2005).

– *Détection de relations sémantiques entre concepts* : la détection de relations sémantiques entre concepts est basée sur des techniques statistiques de l'apprentissage supervisé en utilisant des corpus d'apprentissage qui représentent la connaissance biomédicale (Sibanda, 2006). L'analyse lexicale et l'analyse syntaxique y sont intégrées dans le but d'améliorer la précision de la détection de relations entre concepts.

– *Identification de statuts cliniques des patients* : la plupart des systèmes sont supportés par des méthodes d'apprentissage supervisé et d'algorithmes basés sur des règles de production (Uzuner *et al.*, 2008).

2.3. *Notre contribution dans le domaine*

A notre connaissance, il n'existe pas à ce jour de processus et/ou modèle d'indexation déployé spécifiquement pour les DMP. Notre objectif est alors de proposer un tel modèle en utilisant le thésaurus MeSH. Plus précisément, notre contribution porte principalement sur l'indexation sémantique de l'information explicite (et non implicite) contenue dans le texte intégral des documents qui composent le DMP et ce, en proposant :

– une méthode pour la désambiguïsation des descripteurs sémantiques issus de MeSH, en tenant compte de leur contexte local dans le document : comparativement aux autres méthodes de désambigüisation proposées dans la littérature biomédicale, notre méthode est basée sur le contexte local des concepts dans le document en exploitant la hiérarchie sémantique de MeSH pour identifier le sens correct et non les métadonnées comme les termes/qualificatifs.

– un index sémantique construit selon un schéma de pondération qui combine la spécificité des concepts dans les documents et leur centralité dans les dossiers.

3. Indexation des dossiers médicaux des patients : quel processus et modèle d'indexation?

Dans le but de répondre aux spécificités du dossier patient, nous proposons un processus d'indexation qui est l'enchaînement de deux principales étapes illustrées dans la figure 1 : *annotation sémantique* puis *génération de l'index sémantique*.

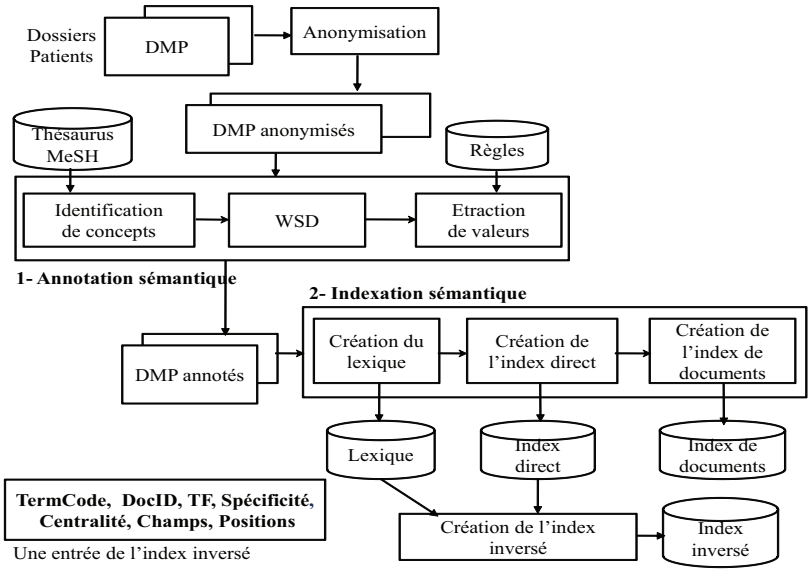

Figure 1. *Processus d'indexation sémantique de DMP*

3.1. *Annotation sémantique*

Cette étape a deux principaux objectifs : a) identifier les concepts associés aux termes extraits de MeSH ainsi que les valeurs associées dans le document, b) désambiguïser les concepts. Avant d'aborder la suite, nous clarifions notre vocabulaire à travers les définitions suivantes :

Définition 1 : Un concept *c* désigne un ensemble de synonymes selon la terminologie de MeSH. Celui-ci se représente par un noeud qui peut appartenir à plusieurs hiérarchies dont chacune correspond à un des 16 domaines : *A-Anatomie*, *B-Organismes*, *C-Maladie ...*

Définition 2 : Un terme *t* est une chaîne de caractères comprenant un mot ou un groupe de mots. Selon la définition de MeSH, à chaque concept est associé un terme préféré qui le représente. Les autres termes synonymes sont dits termes non préférés.

Définition 3 : L'ensemble des sens candidats d'un concept *c*, noté *syn(c)*, est représenté par une hiérarchie de noeuds dans la poly-hiérarchie associée au concept *c*. Chaque hiérarchie est identifiée par un numéro d'arbre (MeSH TreeNumber).

Définition 4 : La relation *is-a* relie les concepts d'une même hiérarchie.

3.1.1. *Identification de concepts et valeurs*

L'identification puis l'annotation conceptuelle des documents sont basées sur l'association des termes de longueur maximale aux entrées de MeSH selon l'algorithme *Left Right Maximum Matching* (Nguyen *et al.,* 2006). A chaque occurrence du concept, on associe une valeur qui peut être qualitative sous forme d'un adjectif (ex : palliatif, aigu, etc.) ou quantitative sous forme de valeur numérique (taux de glucose, pression artérielle, etc.). Les valeurs sont identifiées à l'aide d'expressions régulières correspondant à des patrons d'extraction que nous avons définis.

3.1.2. *Désambiguïsation des concepts*

Un concept de MeSH peut être localisé dans plusieurs hiérarchies et à différents niveaux dans une hiérarchie, correspondant ainsi à différents sens candidats identifiés par des identifiants d'arbres (tree number). Par exemple, le concept "Pain" se trouve dans cinq branches de quatre hiérarchies différentes de MeSH et admet les concepts les plus génériques suivants : *Nervous System Disease* (C10); *Pathological Conditions, Signs and Symptoms* (C23) ; *Psychological Phenomena and Processes* (F02) ; *Musculoskeletal and Neural Physiological Phenomena* (G11). Ceci pose le problème de l'ambiguïté liée à la polysémie que nous résolvons en exploitant le contexte du concept dans le document correspondant. Notre méthode de désambiguïsation est fondée sur : (a) l'hypothèse d'unicité du sens d'un concept dans le document (Gale *et al.,* 1992), (b) la corrélation des sens des concepts voisins : les sens associés à des concepts voisins sur une fenêtre (contexte) sont sémantiquement proches dans le thésaurus, (c) la priorité du sens est définie selon la précédence des concepts : le concept le plus à gauche détermine le sens global de la suite du discours, ce qui crée une chaîne sémantique du discours à partir du début jusqu'à la fin du document.

Nous calculons alors de proche en proche, le sens du concept dans le document par la similarité entre celui-ci et son voisin précédent désambiguïsé. En se basant sur l'hypothèse (a), une fois que le concept est désambiguïsé, son sens est propagé pour toutes ses occurrences dans le document. En considérant la liste de concepts de taille n, $L_n = (c_1, c_2, ..., c_n)$, nous proposons la formule suivante pour identifier le sens optimal du concept c_k :

$$\begin{cases} (s_1, s_2) = \arg\max(\sum_{s_1 \in syn(c_1), s_2 \in syn(c_2)} sim(s_1, s_2)), & \text{k} \leq 2 \\ s_k = \arg\max(\sum_{s \in syn(c_k)} sim(s_{k-1}, s), & \text{k} > 2 \end{cases} \quad [1]$$

où s_k désigne le sens non-ambigu du concept c_k, $syn(c_k)$ l'ensemble de sens candidats du concept c_k et $sim(s_1, s_2)$ la similarité basée sur les hiérarchies de s_1 et s_2.

La similarité entre deux termes est calculée en utilisant une similarité de graphes des hiérarchies de concepts associés selon la formule de (Leacock *et al.*, 1998) :

$$sim(s_1, s_2) = -\log(length(s_1, s_2)/(2 * D)) \quad [2]$$

où $length(s_1, s_2)$ est le chemin le plus court entre s_1 et s_2 et D est le niveau le plus profond dans la hiérarchie.

3.2. *Génération de l'index sémantique*

Notre objectif ici est de générer un index sémantique contenant à la fois les concepts identifiés selon l'approche de désambiguïsation précédente et les termes qui ne correspondent pas à des entrées de MeSH. Nous adoptons les notations suivantes :

– Dossier médical du patient : $DP^{(k)} = \bigcup_{j=1}^{|DP^{(k)}|} D_j$

– Document dans le dossier patient $DP^{(k)}$: $D_j^{(k)}$

– Index sémantique $I = \{c_1, c_2, ..., c_n\}$, où c_i peut être un concept de MeSH ou un terme du vocabulaire, n est le nombre de termes de l'index.

Dans le but de mettre en évidence l'importance des concepts dans les documents d'une part et dans les DMP d'autre part, nous proposons un descripteur à chacun d'eux, défini par un schéma de pondération spécifique.

3.3. *Descripteur du DMP*

L'idée est de représenter le DMP comme un résumé sémantique du profil du patient décrit par les documents le composant. Pour cela, nous mesurons l'importance d'un concept par sa centralité avec ceux de l'index sémantique. Formellement :

$$\begin{array}{l} < DP^{(k)} >:< w_1^{(k)}, w_2^{(k)}, ..., w_n^{(k)} > \\ w_i^{(k)} = (1 + g(c_i))\left|DP^{(k)}\right| * \sum_{D_j \in DP^{(k)}} TF(c_i, D_j) * IDF(c_i) \end{array} \quad [3]$$

où $TF(c_i, D_j)$: fréquence du concept c_i dans le document D_j, $g(c_i)$: centralité normalisée de c_i dans la hiérarchie de concepts, avec:

$$g(c_i) = \left\{ \begin{array}{ll} \frac{Deg(c_i)}{MaxRel} & \text{si } c_i \; est \; un \; concept \; de \; MeSH \\ 0 & \text{sinon} \end{array} \right\} \quad [4]$$

$Deg(c_i)$: degré du concept correspondant au nombre de relations avec les autres concepts, $MaxRel$: nombre maximal de relations possède un concept de MeSH. Plus la centralité d'un concept est élevée, plus le document le contenant inclut des informations synthétiques qui sont potentiellement utiles pour représenter le profil du patient.

3.4. *Descripteur du document*

Dans le cas d'un document, notre objectif est de mieux pondérer les concepts précisant le sujet du document selon leur niveau de spécificité dans la hiérarchie. Formellement:

$$\begin{array}{l} < D_j >:< w_{1j}, w_{2j}, ..., w_{nj} > \\ w_{ij} = (1 + h(c_i)) * TF(c_i, D_j) * IDF(c_i) \end{array} \quad [5]$$

où $h(c_i)$: spécificité normalisée du concept c_i, avec:

$$h(c_i) = \left\{ \begin{array}{ll} \frac{Niveau(c_i)}{MaxDepth} & \text{si } c_i \, est \, un \, concept \, de \, MeSH \\ 0 & \text{sinon} \end{array} \right\} \quad [6]$$

$Niveau(c_i)$: niveau du concept c_i, $MaxDepth$: niveau maximal des concepts.
Un document est spécifique s'il contient des informations spécifiques étant exprimées par des termes utilisés. La spécificité du terme est traduit par son niveau de profondeur dans le thésaurus. Par conséquent, le niveau du concept peut être considéré dans la pondération afin d'introduire sa spécificité dans la description du document.

4. Evaluation expérimentale et résultats

Nous avons mené une série d'expérimentations dans le but de montrer l'intérêt : (a) de l'indexation sémantique, (b) de la méthode de WSD. Nous décrivons dans ce qui suit le cadre d'évaluation puis présentons et discutons les résultats obtenus.

4.1. *Cadre d'évaluation*

– *Collection de documents* : comprend un corpus de DMP sur deux années (2008 et 2009) extraits et anonymisés à partir de la base de données de patients de l'Institut Claudius Régaud (ICR [5]). Les caractéristiques de la collection sont présentées sur le tableau 1 :

Nombre de dossiers patients	1495
Nombre de documents indexés	14941
Taille du vocabulaire	27466
Taille moyenne d'un document	168.58
Nombre de requêtes	25
Longueur moyenne de requêtes	3

Table 1. *Statistiques de la collection test*

5. http://www.claudiusregaud.fr/

– *Collection de requêtes test* : nous supposons que la conclusion du compte-rendu résume le statut sanitaire du patient et traduit un potentiel besoin en information médicale pour les médecins. Pour cette raison, un ensemble de 25 requêtes test a été généré automatiquement à partir des textes associés aux conclusions de 25 documents retenus aléatoirement. Pour des fins expérimentales, nous avons généré, pour chaque requête, deux expressions : (1) basée sur les termes et (2) basée sur les concepts.

– *Mesure d'évaluation* : pour évaluer la qualité des index sémantiques générés selon notre approche, nous procédons à l'évaluation de l'efficacité de la recherche basée sur ces index. Pour cela, nous évaluons pour chaque requête la capacité du système à retourner le document pertinent à partir duquel est extraite la requête (au niveau de la conclusion). Pour ce type d'évaluation, nous avons retenu la mesure appropriée MRR (Mean Reciprocal Rank) (Voorhees, 1999) définie comme suit :

$$MRR = \frac{\sum_{i=1}^{n} rang^{-1}(D_i)}{n} \quad [7]$$

4.2. *Résultats expérimentaux*

Pour évaluer notre modèle d'indexation, deux ensembles d'expérimentations ont été menés : le premier en utilisant une indexation classique basée sur les termes simples déployée en utilisant la plateforme Terrier (http://ir.dcs.gla.ac.uk/terrier/) avec la configuration de pondération de référence BM25 (Robertson *et al.*, 2000); nous avons utilisé ce cadre comme la référence (baseline), noté Ind_Clas sur le corpus initial puis sur le corpus annoté, noté Ind_Clas_Ann. Le second ensemble est basé sur l'application d'une indexation sémantique aux documents et requêtes comme décrit précédemment. Pour évaluer l'impact de la désambigüisation, nous avons comparé la configuration d'une absence de désambigüisation en sélectionnant de manière naïve la première hiérarchie de sens associée au concept, comme présentée dans MeSH, notée Ind_Sem_-Des. On a alors comparé les résultats associés avec ceux obtenus en appliquant notre méthode de désambigüisation, notée Ind_Sem_+Des. La figure 2 présente les résultats obtenus pour les 25 requêtes test, calculés par la formule 7. Les résultats obtenus montrent une amélioration significative (15.67%) de l'approche d'indexation sémantique par rapport à l'approche d'indexation classique qui ne prend en considération que des mots séparés. Ceci montre clairement à la fois l'intérêt de l'indexation sémantique et du schéma de pondération proposé.

La comparaison des résultats obtenus selon les scénarios sans désambigüisation et avec désambigüisation montre un taux d'accroissement des performances de 17%. Ceci montre que les contextes locaux extraits à partir des documents du DMP ont permis de mieux préciser leur sens et donc à les sélectionner.

Une analyse plus fine des résultats au niveau requête montre toutefois que les taux d'accroissement sont variables selon les requêtes, ce qui dans notre cas, est lié à la sémantique portée par les documents source qui ont servi à générer ces mêmes re-

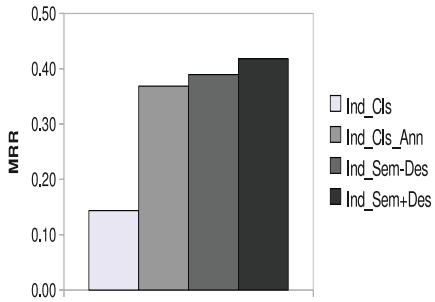

Figure 2. *Résultats d'évaluation*

quêtes. Les raisons de cette variabilité de performance peuvent être expliquées par la variabilité des tailles des index sémantiques associés.

5. Conclusion

Dans ce travail, nous avons proposé et évalué un modèle d'indexation sémantique dédié aux DMP. Le processus d'indexation comprend deux étapes : (1) annotation sémantique qui intègre essentiellement une méthode de désambigüisation de concepts et (2) génération de l'index sémantique qui intègre un schéma de pondération des concepts selon leur spécificité et/ou centralité. L'évaluation de la méthode d'indexation sémantique proposée sur un corpus de DMP a montré que les résultats sont prometteurs. Dans nos futurs travaux, nous envisageons de procéder à une évaluation qui utilise un corpus de dossiers annotés manuellement par les médecins même si cette tâche est couteuse en temps. On envisage également de montrer l'intérêt précis de notre schéma de pondération en le déployant sur des corpus de test avec des configurations différentes de DMP comme celui fourni par I2B2 qui ont la spécificité de contenir des résumés de situations de patients.

Remerciements *Nous tenons à remercier l'Institut Claudius Régaud qui nous a fourni le cadre ayant permis de réaliser ce travail.*

6. References

Andreopoulos B., Alexopoulou D., Schroeder M., « Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering », *Int. J. Data Min. Bioinformatics*, vol. 2, n° 3, p. 193-215, 2008.

Aronson A. R., « Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program », *Proc. AMIA Symp.*, 2001.

Avillach P., Joubert M., Fieschi M., « A Model for Indexing Medical Documents Combining Statistical and Symbolic Knowledge », *Proc. AMIA Symp.*, 2007.

Cai L., Hofmann T., « Hierarchical document categorization with support vector machines », *Proc. CIKM'04*, p. 78-87, 2004.

Friedman C., « Semantic Text Parsing for Patient Records », *USA: Springer*, 2005.

Friedman C., Alderson P. O., Austin J. H., Cimino J. J., Johnson S. B., « A general natural-language text processor for clinical radiology. », *JAMIA*, vol. 1, n° 2, p. 161-174, 1994.

Gale W. A., Church K. W., Yarowsky D., « One sense per discourse », *HLT '91: Proceedings of the workshop on Speech and Natural Language*, p. 233-237, 1992.

Gaudan S., Kirsch H., Rebholz-Schuhmann D., « Resolving abbreviations to their senses in Medline », *Bioinformatics*, vol. 21, n° 18, p. 3658-3664, 2005.

Hersh W., *Information Retrieval: A Health and Biomedical Perspective (Health Informatics)*, 2008.

Kim W., Aronson A. R., Wilbur W. J., « Automatic MeSH term assignment and quality assessment », *Proc. AMIA Symp.*, 2001.

Kolářik C., Hofmann-Apitius M., Zimmermann M., Fluck J., « Identification of new drug classification terms in textual resources », *Bioinformatics*, vol. 23, n° 13, p. 264-272, 2007.

Leacock C., Chodorow M., « Combining local context with WordNet similarity for word sense identification », *WordNet: A Lexical Reference System and its Application*, 1998.

Liu H., Hu Z.-Z., Zhang J., Wu C., « BioThesaurus: a web-based thesaurus of protein and gene names », *Bioinformatics*, n° 1, p. 103-105, January, 2006.

Long W., « Extracting diseases from discharge summaries », *Proc. AMIA Symp.*, 2007.

Névéol A., Rogozan A., Darmoni S., « Automatic indexing of online health resources for a French quality controlled gateway », *Inf. Process. Manage.*, vol. 42, n° 3, p. 695-709, 2006.

Névéol A., Shooshan S. E., Humphrey S. M., *et al.*, « Multiple Approaches to Fine-Grained Indexing of the Biomedical Literature », *Pacific Symp. on Biocomputing*, p. 292-303, 2007.

Nguyen V., Nguyen K., Nguyen H., « Word Segmentation for Vietnamese Text Categorization: An online corpus approach », *RIVF06*, 2006.

Pereira S., Neveol A., *et al.*, « Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue », *Proc. AMIA Symp.*, 2008.

Price S. L., Hersh W. R., Olson D. D., Embi P. J., « Smartquery: Context-sensitive links to medical knowledge sources from the electronic patient record », *Proc. AMIA Symp.*, 2002.

Robertson S. E., Walker S., Beaulieu M., « Experimentation as a way of life: Okapi at TREC », *Information Processing & Management*, p. 95-108, 2000.

Schiemann T., Leser U., *et al.*, « Word Sense Disambiguation in Biomedical Applications: A Machine Learning Approach », *Information Retrieval In Biomedicine*, p. 142-161, 2008.

Sibanda T., Was the patient cured? Understanding semantic categories and their relationships in patient records, PhD thesis, Massachusetts Institute of Technology, 2006.

Subramaniam L. V., Mukherjea S., Kankar P., *et al.*, « Information extraction from biomedical literature: methodology, evaluation and an application », *Proc. CIKM'03*, p. 410-417, 2003.

Uzuner O., Goldstein I., Luo Y., Kohane I., « Identifying patient smoking status from medical discharge records », *JAMIA*, vol. 15, n° 1, p. 14-24, January, 2008.

Voorhees E. M., « The TREC8 Question Answering Report », *Proc of TREC8*, p. 77-82, 1999.

Widdows D., Peters S., *et al.*, « Unsupervised Monolingual and Bilingual Word-Sense Disambiguation of Medical Documents using UMLS », *ACL'03 Workshop*, p. 9-16, 2003.